Journal of Autonomous Intelligence (2024) Volume 7 Issue 2
doi: 10.32629/jai.v7i2.797

ORIGINAL RESEARCH ARTICLE

Hybrid machine learning method for classification and recommendation of vector-borne disease

Salim Gulab Shaikh[1,2,*], Billakurthi Suresh Kumar[3], Geetika Narang[4], Nishant Nilkanth Pachpor[5]

[1] *Department of Computer Engineering, Amity School of Engineering and Technology, Amity University, Rajasthan, Jaipur 302006, India*

[2] *Department of Computer Engineering, AIKTC, Kalsekar Technical Campus, New Panvel, Mumbai 410206, India*

[3] *Department of Computer Science and Engineering, School of Computer Science and Engineering, Sanjay Ghodawat University, Kolhapur 416118, India*

[4] *Department of Computer Science and Engineering, Trinity College of Engineering and Research, Savitribai Phule Pune University, Pune 411048, India*

[5] *Department of Computer Science, Indian Institutes of Management and Studies Savitribai Phule Pune University, Pune 411048, India*

*** Corresponding author:** Salim Gulab Shaikh, shaikhsg2@gmail.com

ABSTRACT

Vector-borne diseases (VBD) are a class of infectious illnesses that are transmitted to humans and animals through the bites of arthropod vectors, such as mosquitoes, ticks, and fleas. These diseases are caused by a variety of pathogens, including bacteria, viruses, and parasites, and are a significant global public health concern. Vector-borne diseases are prevalent in many parts of the world, particularly in tropical and subtropical regions, where the vectors thrive. This research has contributed by constructing a hybrid machine learning based prediction model, which helps to discover patients who are infected by vector-borne disease at an earlier stage and also helps with the categorization and diagnosis of severe vector-borne disease. The model that has been proposed is made up of units: data conversion, data preprocessing, normalization, extraction of feature, splitting of dataset, and classification and prediction unit. The fact that the suggested prediction model is capable of identifying vector-borne disease in its early phases as well as categorizing the kind of disease using the medical report of a sufferer is one of the innovative aspects of the model. The 7 distinct conventional machine learning and single hybrid machine learning (HML) are applied for classification and Recurrent Neural Network (RNN) based reinforcement learning are utilized for recommendation. In order to evaluate the effectiveness of the system that's been proposed, a number of tests were carried out. A dataset consisting of 1539 different cases of a disease transmitted by vectors has been collected. The 11 common vector-borne diseases namely malaria, dengue, Japanese encephalitis, kala-azar and chikungunya were taken for experimental evaluation. The performance accuracy of the proposed prediction model has been measured at 98.76%, which assists the healthcare team in making decisions on a timely basis and ultimately helps to save the patient's lives. The final phase system provides the recommendation for those classifiers resulting in four different classes such as normal, mild, moderate and severe respectively. The recommendation is also demonstrating future direction for cure of vector borne disease.

Keywords: vector-borne disease prediction; natural language processing; feature extraction and selection; machine learning; hybrid machine learning

1. Introduction

The advancement of science and technology has resulted in an improvement in people's quality of life[1]. However, there are still many obstacles to overcome in order to safeguard human life from the onslaught of a wide variety of diseases that are also advancing at the same rate as life itself. In the wake of the development of artificial intelligence (AI) and machine learning (ML), smart machines have also been developed[2]. The

ARTICLE INFO

Received: 14 July 2023
Accepted: 19 September 2023
Available online: 25 December 2023

machine learning algorithms are proposed for vector borne disease outbreak using various feature extraction and selection techniques[3,4]. Moreover, the feature extraction and selection with optimization algorithms such as Particle Swarm Optimization (PSO) with Support Vector Machine (SVM) for disease detection and classification[5]. These infectious diseases can be transmitted from person to person, from animals to people, or from any other element in the environment with VBD detection in high resolution based next generation[6–9]. Patients in rural regions are the ones who are most negatively impacted by this dearth of medical staff. This is due to the fact that most medical professionals are unwilling to work in rural areas, which is where the majority of the Indian population lives. A prediction model that is based on machine learning can be a helpful tool that can assist healthcare staff as well as people who are seeking treatment. In addition, there remains a significant obstacle in India regarding the detection, prevention, and treatment of an outbreak of illnesses that are transmitted by vectors. Because of these factors, it is vital to have a thorough plan in order to ensure that increased or emerging viral activity is discovered prior to the occurrence of an outbreak[10–12].

In most applications machine learning based predictive models for detection and classification of dengue disease is more effective in real time scenarios[13]. However mobile medical assistants are also supplied by health care departments in various countries for cure, prevention and treatment of such diseases[14]. With the use of science and technology, people have been able to establish a clinical diagnosis system over the course of time that enables them to recognize and treat a variety of ailments. The field of medical science has contributed significantly to the prevention and treatment of a wide variety of illnesses through the development of numerous diagnostic procedures and medical tests[15,16]. These machines are capable of consistently keeping an eye on a medical illness as well as the health parameters being tracked by the machine in a variety of medical settings. Nonetheless, the research field of vector-borne diseases is still in its initial stages, and there is a requirement for in-depth studies in a variety of areas that are related to vector-borne diseases[17–19]. These areas include the investigation of various vectors that affect human beings and animals, the preventing of different types of vectors, the study of environmental factors in a wide range of geographical places, a test case of the diagnosis and therapy of these illnesses, and the development of new models for earlier detection of these disease using machine learning techniques[20].

Many people around the world are suffering with diseases that are transmitted by vectors[21–23]. Every single nation on the face of the earth is impacted by these diseases on a yearly basis. These disorders have become a significant burden for the countries that are still developing. Seasonal infections that are transmitted by vectors, such as chikungunya, influenza, dengue, and kala-azar, are those that tend to spread owing to shifts in the natural surroundings and shifts in geographic location. In India, vector-borne illnesses are rapidly expanding throughout the country in a variety of regions. There is a significant amount of difficulty involved in regulating and avoiding the occurrence of diseases of this kind because of the wide variety of environments, climates, and geographical settings. Plasmodium, chikungunya, dengue, and kala-azar are the most significant diseases that are transmitted by vectors and are among the most prevalent diseases in India. In India, there is a significant obstacle to overcome in terms of preventing and managing these diseases. It is now essential to find a remedy to these breakouts by making use of the most recent technological developments, such as information systems and machine learning[24].

The purpose of this research is to evaluate the indications of vector-borne disease and investigate the effect of clinical test factors on those symptoms. The creation of a prediction model for vector-borne illnesses in India through the application of machine learning strategies is the primary purpose of this research project. It has been suggested that the following objectives be accomplished in order to complete this work successfully.

- To determine several important characteristics that are associated with diseases transmitted by vectors.
- To acquire and initially preprocess both primary and secondary data.
- To suggest and develop a model for the prediction of diseases transmitted by vectors.
- To test and evaluate the performance of the proposed model.

The paper is divided into several parts. In part II, related work of various machine learning prediction models of vector-borne disease is described in detail. In part III, the framework of proposed hybrid machine learning based prediction model of vector-borne disease is discussed in detail, in part IV, various experimental findings and performance comparison of proposed HML with conventional machine learning techniques is illustrated. Finally, in part V, proposed methodology is concluded.

2. Literature survey

According to Karn et al.[1] it is possible to become infected with contagious diseases as well as vector-borne diseases. These illnesses have extremely similar symptoms, the majority of which manifest themselves a few days after infection. The accurate identification of these disorders is now made possible with the assistance of modern technology. An accurate diagnosis is required to make sure that approach and prescription drugs are administered, which necessitates the need for an automation process to anticipate possible infectious diseases. In order to make sure that appropriate authorization and medicines are prescribed, an early diagnosis is essential. This necessitates the development of a method that provides the patient with the ability to differentiate between these illnesses and identify the possible illness based on the person’s illness. Once a diagnosis of the disease has been made, the next step is to determine the most effective course of treatment depending on the anticipated form of the condition. The application of this method for clinical diagnosis is performed with the assistance of artificial neural networks (ANNs) the training of which is accomplished through the use of the back-propagation algorithm. When compared to the rule-based model, the accuracy of the system increases when ANNs are utilized in medical diagnosis. Furthermore, when combined with the back propagation algorithm and the gradient optimization method, the findings are even more accurate.

According to research conducted by Kaur et al.[2] among the many dangers that confront our planet today, diseases transmitted by vectors pose the biggest risk. Arboviruses get a longstanding experience of causing disease; but, in recent years, they have become increasingly common and are impacting greater populations. This is despite the fact that they have been around for a very long time. This is attributable to a number of factors, including a rise in the amount of people who travel by air and uncontrollable numbers of mosquito vectors. It may be possible to use ML and neural networks in an effort to limit the spread of epidemics caused by lethal infectious diseases. The methods, data, and quality measurements that are utilized in applications for forecasting and diagnosing fatal infectious diseases were not discussed in a number of the research that were conducted. This article presents a summary of research on two primary approaches that have been shown to be effective in preventing the spread of lethal disease outbreaks. This study investigates the present advancements, challenges, and potential future applications of ML and DL to identify and anticipate fatal disease epidemics in order to cut down on the likelihood that an infection will be transferred to more people. In this study, we investigate the techniques, datasets, parameters, and performance indicators of previously conducted research.

Raizada et al.[3] concentrated on precise estimates of the emergence of vector-borne diseases over the Indian-subcontinent for the diseases chikungunya, malaria, and dengue. The developed framework has been scrutinized and improved using data gathered all throughout India between the years 2013 and 2017. It has presented a Convolutional Neural Network (CNN) approach for the prediction of outbreak risk that makes use of contrasting data. To the best of our knowledge, none of the earlier works have focused on comparing data in the field of medical data analysis. The proposed CNN algorithm has an accuracy of prediction that is 88%.

Davi et al.[4] describe a machine learning strategy for predicting the severity of dengue fever simply using data from human genomes. This approach is based on the assumption that human genomes contain all of the relevant information. The genetic makeup of 102 Brazilian dengue patients and 102 healthy controls was analyzed in order to genotype 322 innate immune system single nucleotide polymorphisms. A support vector machine approach is used in the suggested model in order to determine the optimal locus classification subset. Following this, an ANN is utilized in order to categorize patients as having dengue fever or dengue hemorrhagic fever. Training the ANN on 13 critical immunological Single nucleotide polymorphisms (SNPs) that were either dominant or recessive generated median rates of accuracy higher than 86%, as well as specificity and sensitivity values that were above 98% and 51%, correspondingly. Even in situations where the person is not infected with dengue, the suggested classification system, which uses only genomic markers, can be employed to detect those who have a high risk of acquiring a severe form of the disease caused by dengue. Based on the findings, it appears that the genetic background has a significant role in determining the dengue phenotype. The method that has been presented in this article can be extended to additional diseases that are based on Mendelian inheritance or are genetically impacted.

Singh[5] proposes a system of particle swarm optimization to support vector machine based diagnostic systems as a method for early dengue diagnosis. The primary purpose of this research is to evaluate how successful the PSO and SVM are when it comes to extracting the dengue dataset. The precision of the machine-learning algorithm is the goal of this research, which aims to improve it. The recommended methodology was validated even further by employing a number of different standard dengue categorization data sets. An assessment is conducted between the suggested method and the one that is currently being used by looking at how well each approach meets the various standard service quality standards. The findings of the experiments indicate that the strategy that was advised is the more effective option.

Saturi[6] constructed a forecast model to manage the outbreak of dengue sickness. This model provides medical experts the ability to design, plan, and handle the disease at a very early stage in its progression. In addition, research was conducted on the enhancement of a variety of approaches for evaluating and predictive modeling through the utilization of measurable, numerical analysis of machine learning. Exploration of data sources, analysis of data sources, methods for data preprocessing, statistical modeling, dengue forecasting models, and assessment methodologies are the six primary problems that need to be resolved in order to determine the cause of dengue sickness. Traditional approaches have a number of drawbacks, one of the most significant of which is that in order to enhance dynamic properties, these methods require a significant amount of data to be processed. Based on the evaluation of the various current methods, it is possible to claim unequivocally that the k-means clustering method using a fuzzy based system has excellent accuracy and that it considerably improves the analysis and prediction of dengue sickness. This is the case. The patient records associated with dengue sickness are clustered using the k-means technique and divided into k subgroups. The k-means clustering technique helps improve the assessment or forecast of dengue sickness since the dengue dataset was fully clustered. In the same vein, the fuzzy-based system, the input elements, and converting these informative variables into fuzzy functions of membership will result in improved decision-making regarding the dengue forecasting model. As a result, the problems that have been identified as a result of extensive research offer a helpful foundation for epidemiological and public biomedical research.

Sarder et al.[7], climate data provide an accurate basis for their dengue epidemic prediction. A dataset on dengue that was compiled by the Meteorology Department and the Directorate General of Health Services (DGHS), Bangladesh. It includes information on climatic parameters as well as dengue cases that occurred between 2019 and 2021. The entire dataset is divided up into 70:30 portions, with the first portion serving as training and the second portion serving as testing. The SVM, random forest, decision tree, logistic regression, naive bayes, AdaBoost and gradient boost classifier are some of the supervised machine learning techniques that we use to make such predictions of accuracy. Last but not least, among these algorithms, the SVM has the best accuracy, coming in at 96.73%.

The method proposed by Ghaffari et al.[8] makes use of data from Twitter and applies machine learning methods at several spatial scales in order to circumvent the limits of the platform and produce results with the appropriate level of detail. Validation of the suggested method was performed using the Zika epidemic that occurred in Florida in 2016. The most important aspect of this study is the proposal of an innovative method that makes use of machine learning techniques applied to data collected from social media platforms in order to determine the potential threat posed by the introduction of vector-borne diseases at a spatial resolution that is fine enough to enable efficient invasion. It will pave the way for a new generation of models for assessing the risk of epidemics and has the potential to radically improve public health by pinpointing specific areas in need of targeted intervention.

Siddiq et al.[9] aim to identify which of four models, one linear (linear regression) and three nonlinear (SVM, RF, DT), had the most accurate ability to forecast data fragmentation and transmission. The models for making predictions were developed using data from verified dengue fever cases that were found in Jeddah city in Saudi Arabia, in addition to the humidity and temperature which are the two factors that have the most link to known instances. The SVC prototype has the best performance among the models that were tested. It achieved a predictive performance of 76%, while the regression model, the RF regression model, and the DT regression model achieved a predictive performance of 52%, 55%, and 57%, respectively.

The purpose of this study by Akramin Kamarudin et al.[10] is to examine the current framework for the prediction of MBD outbreaks and to propose an improved version of the framework that includes an entomological index component. A fresh conceptual approach is presented here, one that makes use of machine learning to boost future MBD epidemic predictions.

CNN models are suggested to be utilized by Ahmed et al.[11] in order to automatically diagnose malaria based on photographs of red blood cells. After applying DenseNET-121 for extraction of features, the training algorithms included DT, RF, k-nearest neighbor, SVM and AdaBoost classification methods. Among the seven models that have been recommended as a consequence of this research, the efficiency of the DenseNET-121 with XGBoost Classifier paradigm is the greatest at 96.3%.

As per Dhaka et al.[12] it is possible to forecast the weather by making use of the sophisticated equipment and satellites that are now on the market. With the help of this prediction, it should be feasible to foresee who the next victims of the epidemic will be. Four different algorithms, including RF regression, DT regression, support vector and multiple linear regressions, are utilized in the process of putting this epidemic alarm system into action. The system uses state-wise case data from 2013 to 2017 for dengue, but data from 2013 to 2016 is utilized for chikungunya. The information used for dengue are from 2013 to 2017, while the information used for chikungunya are from 2013 to 2016. In the context of dengue, the algorithm was educated using the data from 2013 to 2016, and it has been used to make predictions for the year 2017. On the contrary hand, the system was trained using the data from the years 2013 through 2015, and it has been used to make forecasts for the year 2017 with reference to the chikungunya virus. At long last, a comparative study of the four algorithms that were utilized in the treatment of both disorders has been carried out.

According to research done by Ganapathi Raju et al.[13], dengue is a mosquito-borne viral disease that can be lethal and has lately emerged as an issue on a global scale. Temperature, rainfall, and moisture are the three primary climatic conditions that remain essential for the survival, propagation, and growth of mosquitoes, which can further affect the existence and number of mosquitoes. Temperature is the most important component. Taking into consideration the dengue data from the state of Kerala, a predictive model for the prevalence and incidence of dengue fever has been developed. In the section titled "prediction," the method computes an estimate of the total number of dengue fever cases that could occur during the given month. The association between humidity, heat, precipitation, and dengue cases was investigated through the use of meteorological analysis. Linear regression, support vector regression, and kernel ridge were the three techniques for machine learning that were utilized in order to make the prediction. Both meteorological and geo geographic analysis on the data have been the subject of comparable research studies. For the purpose of the geographic data, the following criteria were taken into account the distance of the neighborhood from the tropics; the shortest path between the district and the nearest seaside; the proportion of the district that is covered in forest; and finally, the altitude of that specific jurisdiction.

Devarajan et al.[14] show how machine learning can be utilized to detect Parkinson's disease (PD) and, as a result, provide early diagnosis utilizing nonclinical data pertaining to the patients. In this study, novel ensembles are constructed to enhance the diagnosis process, and the experimental findings reveal that enhanced versions of ANN classifiers can produce results that are 13.4% more accurate than the classic ANN classifier. The fact that PD affects people all around the world and has a difficult diagnostic process makes it a complicated medical illness to treat. In addition, the early diagnosis of PD is essential to the patient's likelihood of making a full recovery, and any errors in diagnosis might result in an irreparable loss for the patient. Also, as a result of the study, a reliable diagnostic tool for Parkinson's disease has been established. This tool helps to detect the disease in its earliest stages by analyzing individuals' voice data. As a result, better clinical decisions with respect to PD will be able to be made, which will lead to improved health care services.

The research conducted by Jayampathi et al.[16] describes a mobile medical aid and analytical platform for patients diagnosed. The mobile application takes a unique strategy and makes use of the most suitable techniques in order to support the identification of dengue patients through the use of the chatbot, the analysis of skin problems, the analysis of blood findings, and the analysis of the features and functionality of dengue-infected areas. Users who have previously registered for the system are able to verify their dengue status by logging in. Technology such as natural language processing, ANN, ML, image processing, CNN, and Android are utilized throughout the development process. The concept for a mobile application is developed and evaluated, with the goal of conducting further testing and maybe putting the application into production. The findings demonstrate that the methods used to analyze dengue conditions are efficient.

3. Research methodology

A significant obstacle on a global scale is the prevention and management of epidemics of diseases transmitted by vectors. The population, density, and type of mosquitoes that are present in cycle-based seasonally, atmospheric, and geographical settings all play a role in determining how quickly vector-borne diseases spread. Vector infections continue to be a problem in India because to the vast cultural, geographical, and climatic variety that exists there. A certain number of individuals are diagnosed with any of these diseases each year. As a direct consequence of this, the vast majority of them succumbed to the diseases because they were either not detected, detected too late, or treated too late. The biting of an infected mosquito is the most common way that diseases are spread that are carried by vectors like mosquitoes. The bite of a mosquito can spread a number of different diseases, including plasmodium, zika, yellow fever, dengue, Nile. There are a number of diseases that are transmitted by vectors that can be found in India. It is abundantly obvious from the body of published research that the majority of the existing architecture has been created in the field of diseases transmitted by vectors. However, the researchers have not yet been successful in developing a general model for diseases transmitted by vectors. In India, research on a few specific diseases transmitted by vectors is still in its early stages. Therefore, there is a significant requirement for the development of a general prediction model for diseases of this nature. As a result of this, efforts have been made to construct a prediction model employing hybrid machine learning for diseases that are transmitted by vectors.

The proposed approach for prediction is made up of five different units: data transformation, data pre-processing and feature extraction, splitting of dataset, classification and prediction unit. Each unit is responsible for a distinct set of tasks, and the result of one unit serves as the input for the subsequent unit (as depicted in **Figure 1**).

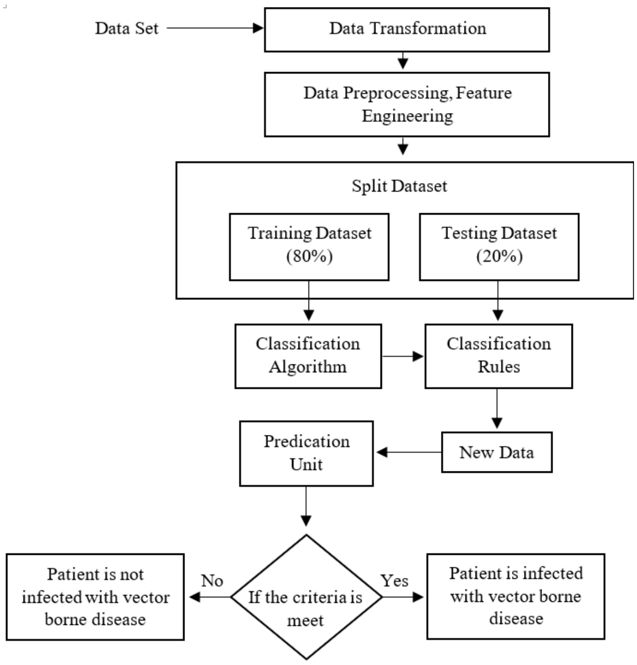

Figure 1. Proposed HML based prediction model for vector-borne disease.

The data transformation is the first unit. The primary purpose of this unit is to retrieve the information from the repository and convert the unstructured or semi-structured data into relational form. In order to accomplish this goal, the required data was obtained from the data repository in the format of SQL statements or documents in the csv file format. The second unit is data preprocessing and feature extraction and its purpose is to clean the data, scale the data, and normalize the data such that it meets the requirements of the suggested prediction model. The primary responsibilities of this module include the production of scales for each of the features relating to diseases transmitted by vectors, the cleansing of data, and the management of outliers in the dataset. In order to accomplish this objective, a step-by-step method is carried out to produce a dataset for the prediction model that is of the highest possible quality and is devoid of errors. The splitting of datasets is the third unit, and in this unit, a dataset is divided into two categories: the training and the testing dataset, with a ratio of 80:20. The effectiveness of the dataset is evaluated with the help of the testing dataset, while the training data set is employed for the process of training the prediction model. The purpose of the training of a prediction model using the training dataset and the test dataset is accomplished by the classification unit, which is the fourth unit of the overall process. The fifth and final unit of the system is the prediction unit, and its purpose is to enable the user to make forecasts based on data collected from the surrounding environment.

Early disease detection as well as disease type categorization are both possible with this component. The operation of each unit of the model that's been proposed for predicting diseases transmitted by vectors has been described in detail.

Data transformation: the data transformation is a process of transforming raw and semi-structured data into a state suitable for data analysis. The data transformation unit consists of 8 stages. In the first stage, the data are fed into a panda dataset in the form of semi-structured data utilizing a variety of repositories. After that, the data is saved in python. In the second stage, the procedure entails deleting unnecessary as well as repetitive entries from the dataset. The third stage of the procedure involves converting string data into textual data and saving it in the appropriate column of the dataset. After going through this process, the data that is accessible at that point can be utilized for data analysis. In stage 4, the translation process involves transforming the category data into numerical form. This procedure is carried out with the use of a scale that is assigned to a specific characteristic. The suggested prediction for every feature has its own level for storing categorical data. For example, the scale for fever includes the categories high, medium, low, and null. In stage 5, the procedure of transforming the data from the time stamps into the appropriate format is completed. In step 6, the data that has been refined are brought up to date and saved in the dataset so that it can be preprocessed and normalized.

Data pre-processing and feature engineering: this section is extremely important to the overall prediction model. This part is responsible for two significant responsibilities, namely, the preprocessing of the data and the feature engineering.

Data preprocessing: at this stage, the job of processing the data and cleansing the data is accomplished. The data preprocesses into 8 stages. The dataset is imported from the transformation unit as the first stage in the process. The second stage involves selecting the most significant attributes and then retrieving the data from the database. In the third stage, the investigator located the data that were missing from the dataset and replaced it with the appropriate data values. In the fourth stage, it purges the dataset of any records that are a duplicate of others. In the fifth phase, it locates the values that are beyond the normal range and then removes those values from the dataset. The job of normalizing is carried out on each characteristic of the dataset at the sixth stage of the process. In the eight stage, all of the activities that are completed inside the scope of this unit have their server-side versions updated and saved.

1) **Deal with missing values:** decide whether to impute missing data or remove rows/columns with missing values.
2) **Handle duplicate values:** identify and remove duplicate records.
3) **Correct inconsistent data:** standardize or correct data that is inconsistent or incorrectly formatted.
4) **Data Scaling:** normalize or standardize numerical features to ensure they have similar scales (e.g., using Min-Max scaling or Z-score scaling).
5) **Encoding categorical variables:** convert categorical variables into numerical format (e.g., one-hot encoding or label encoding).
6) **Handling skewed data:** apply transformations like log or square root to reduce skewness in the data.
7) **Feature engineering:** create new features based on domain knowledge or data patterns.
8) **Splitting of data:** in this unit, a sizable dataset is partitioned into a training dataset and a test dataset in order to construct a prediction model for diseases transmitted by vectors. The dataset has been divided into two portions with a 70:30 split between them.

Feature extraction and selection: feature extraction and feature selection are essential processes in machine learning and data analysis, especially when dealing with numeric data. They help improve model performance, reduce overfitting, and enhance interpretability. It is a process in machine learning and data analysis that involves choosing a subset of the most relevant and important features (variables or columns) from a larger set of available features in your dataset. The goal of feature selection is to improve model performance, reduce overfitting, enhance computational efficiency, and increase interpretability by focusing on the most informative attributes while discarding irrelevant or redundant information. The TF-IDF, WTF, co-relation coefficient and autoencoder based features are extracted during the execution for robust module training.

Classification: classification is a type of predictive modeling task in which a ML model is trained using a training dataset and then uses that model to classify some event or object depending on the input that is provided. Classification is a type of supervised learning function that involves developing a prediction model with the help of a training set. A prediction model's efficacy can be evaluated using the test dataset as a basis for comparison. The primary goal of classification is to make an accurate prediction of the class label that should be assigned to each input.

Recommendation module: the evaluation of symptoms as a basis for the early diagnosis of diseases transmitted by vectors is the primary task of this section. On the basis of the medical data, the second purpose of this unit is to determine if the disease transmitted by vectors is mild or severe. This determination will be made based on the classification of the disease. The data of the patient's record is taken from the smartphone app, the web application, and the software applications throughout the entirety of the procedure that comprises the prediction module. Research on data cleaning as well as preprocessing is carried out after the medical record of the patient has been obtained. The record of the patient is next examined to see if it merely contains symptoms or whether it also contains symptoms along with a clinical report. This prediction unit will be utilized for the early stage diagnosis of vector-borne disease whenever the patient's record is obtained with only details relating to symptoms. In addition, if the patient's data is received along with symptoms and a medical report, then this unit is regarded as having confirmed cases.

4. Algorithm design

In the proposed implementation we used 7 conventional machines learning and 1 hybrid machine learning algorithm used for classification. The HML refers to the combination of different machine learning techniques or algorithms to solve a particular problem. This can include a combination of supervised and unsupervised learning methods, as well as other techniques such as rule-based systems or deep learning. When it comes to classification tasks, hybrid machine learning approaches can be particularly effective in improving accuracy and robustness.

HML for classification of VBD:

In prosed HML the three different machine learning algorithms are utilized for identification of HML. Each algorithm evaluated using pertained SVM based module and predict the test results. In evaluation mapping each algorithms predicted label has evaluated with actual label and determines the final label when both are equals. In below we describe in details execution of proposed HML algorithms.

Input: Selected training dataset Training_Data[], selected testing dataset Testing_Data[], Threshold qTh

Output: Result set as output with {Label_class, class_weight}

Step 1: The function provided is used to extract and validate the test data from the Test_Data[] array.

$$Test\_Feature[dataset] = \sum_{m=1}^{n} (Attribute\_Set[A[m], \dots, A[n]\ Test\_Data)$$

Step 2: Choose the features from the test_Feature[dataset] attributes set. The following function is used to extract features and build a feature map.

$$Test\_FeatureMap[t, \ldots, n] = \sum_{x=1}^{n} (t) test\_Feature(x)$$

The features used for the pooling layer are denoted by Test_FeatureMap[x]. Test_FeatureMap contains the features chosen by the convolutional layer and passed on to the pooling layer.

Step 3: In order to classify all of the test data in the sense layer, we must first read the whole taring dataset,

$$Train\_Feature(data) = \sum_{m=1}^{n} (Attribute\_Set[A[m], \ldots, A[n]\ Train\_Data)$$

Step 4: Generate the training map using below function from input dataset

$$Train\_FeatureMap[t, \ldots, n] = \sum_{x=1}^{n} (t) train\_Feature(x)$$

The map of the hidden layer, Train_FeatureMap[t], is responsible for creating the feature vectors used to construct the hidden layer. That uses train data to evaluate all test cases.

Step 5: After the feature map is created, we assign a weight to the degree of similarity between each instance in the dense layer and the characteristics of interest in the pooling layer.

$$Gen\_weight = CalcWeight(Test\_FeatureMap || \sum_{i=1}^{n} Train\_FeatureMap[i])$$

Step 6: Evaluate the current weight with desired threshold

$$If\ (Gen\_weight \geq qTh$$

Step 7: $Res\_List.add$ **(**$trainF.class, weight$**)**

Step 8: Repart step 1 till test_data! = null

Step 9: Return $Res\_List[]$

Both the Train_Feature[] and Test_Feature[] need input for testing the classifier when generating a similarity score between two input items. The two characteristics in question reflect the training and testing instances, respectively. The "Th" represents the denominator used in the selection process of each epoch layer outcome. The notation T[*j*] represents the *j*-th attribute of a testing instance, whereas T[*k*] represents the *k*-th property of a training instance. The feature selection approach is used to extract relevant characteristics from both instances. These selected features are then passed on to the similarity measurement function, as outlined in step 5. The dense optimized results obtained by HML are determined by the number of validated occurrences based on a specified threshold.

5. Result and discussion

In order to evaluate the effectiveness of the system that's been proposed, a number of tests were carried out. A dataset consisting of 1500 different test cases of a disease transmitted by vectors has been collected. During the process of result analysis, the degree of accuracy with which various diseases might be forecasted using the suggested approach has been observed. The dataset has been taken from Lad[24] for training and testing data to accurately detect VBD. We used all attributes including prognosis as class label. The dataset contains 65 attributes including class labels with 5050 records with 11 different class labels. The dataset has been distributed using different cross validations such as 5-fold, 10-fold and 15-fold cross validation.

The entire dataset contains 11 disease types which all are associated with VBD. Consider the below **Table 1**, which illustrates the analysis of vector-borne disease of 1539 different cases with each class information with count details.

Table 1. Data description with No. of patients associated with specific VBD.

Disease name	No. of records
Zika	137
Chikungunya	140
Dengue	127
Japanese_encephalitis	157
Lyme_disease	155
Malaria	154
Plague	146
Rift_valley_fever	145
Tungiasis	132
West_nile_fever	108
Yellow_fever	138

Consider the following **Table 2**, which depicts the performance comparison of proposed HML and other traditional methods of machine learning using performance measurements namely, precision, F1-score, accuracy and recall.

Table 2. Performance Comparison of Proposed HML and Traditional method of ML classifiers.

Algorithms	Accuracy	Precision	Recall	F1-score
ANN	73.03	75.60	74.30	76.10
NB	97.33	96.50	96.40	97.45
J48	90.65	91.30	91.48	92.60
AdaBoost	18.77	19.80	21.50	21.80
RF	96.62	95.90	95.93	96.30
SVM	86.74	87.45	88.10	88.60
MLC	82.91	83.40	84.10	85.25
HML	98.76	98.70	98.40	99.10

It is observed from the above **Table 3** that the performance comparison of proposed HML is better as compared to traditional methods of machine learning. Consider the following **Figure 2** which depicts the performance comparison of the proposed HML predicted model of vector-borne disease and other traditional methods of machine learning.

Table 3. Data distribution for training and testing.

	No. of records
Training	3550
Testing	1539

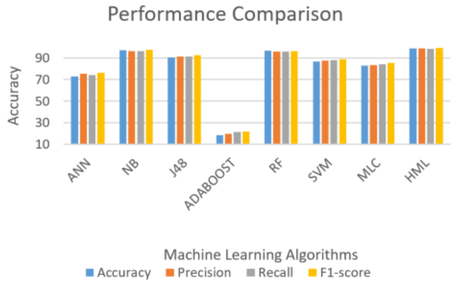

Figure 2. Performance comparison of proposed HML prediction model of vector-borne disease.

It has been observed from the above **Figure 2** that the accuracy of proposed HML prediction model of vector-borne disease is 98.76% which is better than traditional ML classification models namely random forest, decision tree, logistic regression, support vector machine, naive bayes.

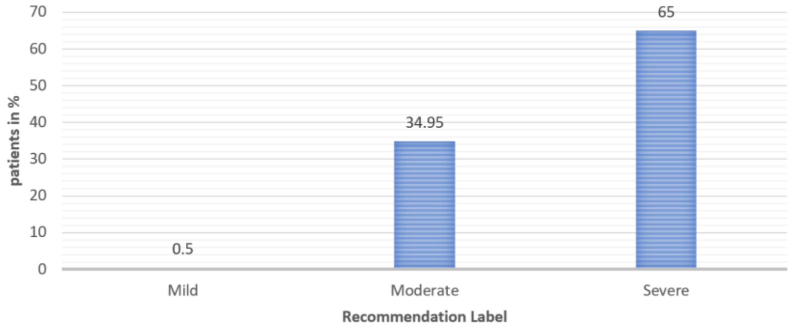

Figure 3. Recommendation using reinforcement learning algorithm for vector bone disease algorithm.

The above **Figure 3** describes a VBD recommendation based on classification results. The 3 different class labels are demonstrated for the entire dataset and it also recommend the future action for mild, moderate and severe profiles. The recurrent neural network (RNN) based feedback approach and reinforcement learning techniques have been used for final recommendation.

6. Conclusion

The system proposed an VBD classification and recommendation using hybrid machine learning techniques. The seven-machine learning single hybrid machine learning algorithms are applied for classification of VBD while RNN based reinforcement learning techniques are utilized for recommendation. The subsequent step involves the extraction of features from both the historical data and the new data. Descriptive statistical analysis is used to investigate the relationships between the various features in order to determine which ones are more highly correlated with one another. The imported dataset is then splitted into two sets, known as the training set and the test set, with a ratio of 80:20 between the two. The proposed prediction model was tested and analyzed with various conventional machine learning algorithms, namely random forest, decision tree, logistic regression, support vector machine, naive bayes. Experiments have been carried out in order to assess how well the suggested HML-based prediction model of vector-borne disease performs. The suggested model was able to achieve an average accuracy of 98.76% after going through the testing and validation process, which is higher than the accuracy attained by other conventional methods of machine learning.

Author contributions

Conceptualization, SGS and BSK; methodology, SGS; software, SGS; validation, SGS, BSK, GN and NNP; formal analysis, SGS; investigation, SGS; resources, SGS; data curation, SGS; writing—original draft preparation, SGS; writing—review and editing, SGS; visualization, SGS; supervision, SGS; project administration, SGS; funding acquisition, BSK. All authors have read and agreed to the published version of the manuscript.

Conflict of interest

The authors declare no conflict of interest.

Abbreviations

Support vector classification—SVC

Hybrid Machine Learning—HML

Support vector Machine—SVM

Random Forest—RF

Decision Tree—DT

Multibacillary Disease—MBD

Structured Query Language—SQL

Term Frequency and Inverse Document Frequency—TF-IDF

Weighted Term Frequency—WTF

References

1. Karn S, Sangole S, Gawde A, Joshi J. Prediction and classification of vector-borne and communicable diseases through artificial neural networks. In: Proceedings of the 2019 International Conference on Intelligent Computing and Control Systems (ICCS); 15–17 May 2019; Madurai, India. pp. 1011–1015.
2. Kaur I, Sandhu AK, Kumar Y. Analyzing and minimizing the effects of vector-borne diseases using machine and deep learning techniques: A systematic review. In: Proceedings of the 2021 Sixth International Conference on Image Information Processing (ICIIP); 26–28 November 2021; Shimla, India. pp. 69–74.
3. Raizada S, Mala S, Shankar A. Vector-borne disease outbreak prediction by machine learning. In: Proceedings of the 2020 International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE); 9–10 October 2020; Bengaluru, India. pp. 213–218.
4. Davi C, Pastor A, Oliveira T, et al. Severe dengue prognosis using human genome data and machine learning. IEEE Transactions on Biomedical Engineering 2019; 66(10): 2861–2868. doi: 10.1109/tbme.2019.2897285
5. Singh KD. Particle swarm optimization assisted support vector machine based diagnostic system for dengue prediction at the early stage. In: Proceedings of the 2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N); 17–18 December 2021; Greater Noida, India. pp. 844–848.
6. Saturi S. Development of prediction and forecasting model for dengue disease using machine learning algorithms. In: Proceedings of the 2020 IEEE International Conference on Distributed Computing, VLSI, Electrical Circuits and Robotics (DISCOVER); 30–31 October 2020; Udupi, India. pp. 6–11.
7. Sarder F, Akter S, Akter S. Predicting dengue outbreak from climate data using machine learning algorithms. In: Proceedings of the 2022 IEEE International Conference on Data Science and Information System (ICDSIS); 29–30 July 2022; Hassan, India. pp. 1–6.
8. Ghaffari M, Srinivasan A, Mubayi A, et al. Next-generation high-resolution vector-borne disease risk assessment. In: Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM); 27–30 August 2019; Vancouver, BC, Canada. pp. 621–624.
9. Siddiq A, Shukla N, Pradhan B. Predicting dengue fever transmission using machine learning methods. In: Proceedings of the 2021 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM); 13–16 December 2021; Singapore, Singapore. pp. 21–26.
10. Akramin Kamarudin AN, Zainol Z, Abu Kassim NF. Forecasting the dengue outbreak using machine learning algorithm: A review. In: Proceedings of the 2021 International Conference of Women in Data Science at Taif University (WiDSTaif); 30–31 March 2021; Taif, Saudi Arabia. pp. 1–5.
11. Ahmed MS, Rahman R, Arefeen ZR, et al. Effort to mitigate malaria via early detection using hybrid machine learning architectures. In: Proceedings of the 2021 31st International Conference on Computer Theory and Applications (ICCTA); 11–13 December 2021; Alexandria, Egypt. pp. 155–159.
12. Dhaka A, Singh P. Comparative analysis of epidemic alert system using machine learning for dengue and chikungunya. In: Proceedings of the 2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence); 29–31 January 2020; Noida, India. pp. 798–804.
13. Ganapathi Raju NV, Krishna PG, Manognya K, et al. Evolution of predictive model for dengue incidence by using machine learning algorithms. In: Proceedings of the 2019 International Conference on Communication and Electronics Systems (ICCES); 17–19 July 2019; Coimbatore, India. pp. 51–59.
14. Devarajan JP, Sreedharan VR, Narayanamurthy G. Decision making in health care diagnosis: evidence from Parkinson's disease via hybrid machine learning. IEEE Transactions on Engineering Management 2023; 70(8): 2719–2731. doi: 10.1109/tem.2021.3096862
15. Nalini C, Shanthakumari R, Venkata PR, et al. Prediction of dengue infection using machine learning. In: Proceedings of the 2022 International Conference on Computer Communication and Informatics (ICCCI); 25–27 January 2022; Coimbatore, India. pp. 1–5.
16. Jayampathi KTK, Jananjaya MAC, Fernando EPC, et al. Mobile medical assistant and analytical system for dengue patients. In: Proceedings of the 2021 3rd International Conference on Advancements in Computing (ICAC); 9–11 December 2021; Colombo, Sri Lanka. pp. 371–376.
17. Feng C, Wu J, Wei H, et al. CRCF: A method of identifying secretory proteins of malaria parasites. IEEE/ACM Transactions on Computational Biology and Bioinformatics 2022; 19(4): 2149–2157. doi:

10.1109/tcbb.2021.3085589

18. Umer M, Sadiq S, Ahmad M, et al. A novel stacked CNN for malarial parasite detection in thin blood smear images. IEEE Access 2020; 8: 93782–93792. doi: 10.1109/access.2020.2994810
19. Prakash N, Balaji VR. Detection of plant disease using swarm intelligence optimization algorithm. In: Proceedings of the 2021 International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA); 8–9 October 2021; Coimbatore, India. pp. 1–5.
20. Mathur N, Asirvadam VS, Dass SC. Spatial-temporal visualization of dengue incidences using gaussian kernel. In: Proceedings of the 2018 International Conference on Intelligent and Advanced System (ICIAS); 13–14 August 2018; Kuala Lumpur, Malaysia. pp. 1–6.
21. Huang LP, Hong MH, Luo CH, et al. A vector mosquitoes classification system based on edge computing and deep learning. In: Proceedings of the 2018 Conference on Technologies and Applications of Artificial Intelligence (TAAI); 30 November 2018–2 December 2018; Taichung, Taiwan. pp. 24–27.
22. Varshini TH, Samatha B. Deep learning technology to identify arboviral disease-dengue prediction. In: Proceedings of the 2022 International Conference on Edge Computing and Applications (ICECAA); 13–15 October 2022; Tamilnadu, India. pp. 1317–1323.
23. Shamim MAR, Anas ABM, Erfan M. Identification of vector and non-vector mosquito species using deep convolutional neural networks with ensemble model. In: Proceedings of the 2022 International Conference on Advancement in Electrical and Electronic Engineering (ICAEEE); 24–26 February 2022; Gazipur, Bangladesh. pp. 1–6.
24. Lad S. Tabular-Vector-Borne-Disease-dataset. Available online: https://www.kaggle.com/datasets/snehalad/tabular-vector-borne-disease-dataset (accessed on 28 April 2023).